

# Convergence of the empirical process in Mallows distance, with an application to bootstrap performance

Richard Samworth<sup>\*†</sup>      Oliver Johnson<sup>\*‡</sup>

February 1, 2008

Running Title: Empirical process in Mallows distance

Keywords: Bootstrap, empirical distribution, empirical process, hazard function, Mallows distance, probability metric, sample mean, Wasserstein distance.

Mathematics Subject Classification: 62E20; 60F25; 62F40.

## Abstract

We study the rate of convergence of the Mallows distance between the empirical distribution of a sample and the underlying population. The surprising feature of our results is that the convergence rate is slower in the discrete case than in the absolutely continuous setting. We show how the hazard function plays a significant role in these calculations. As an application, we recall that the quantity studied provides an upper bound on the distance between the bootstrap distribution of a sample mean and its true sampling distribution. Moreover, the convenient properties of the Mallows metric yield a straightforward lower bound, and therefore a relatively precise description of the asymptotic performance of the bootstrap in this problem.

---

<sup>\*</sup>Statistical Laboratory, DPMMS/CMS, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK. Fax: +44 1223 337956

<sup>†</sup>Email: [rjs57@cam.ac.uk](mailto:rjs57@cam.ac.uk)

<sup>‡</sup>Email: [otj1000@cam.ac.uk](mailto:otj1000@cam.ac.uk)

# 1 Introduction and main results

Different problems in Probability and Statistics naturally lead to different choices of probability metric. One such choice is the Mallows distance, also known as the Wasserstein or Kantorovich distance. This metric has found extensive applications to a wide variety of fields; see Rachev (1984) for a review.

**Definition 1.1** *For  $r \geq 1$ , let  $\mathcal{F}_r$  denote the set of distribution functions  $F$  satisfying  $\int_{-\infty}^{\infty} |x|^r dF(x) < \infty$ . For  $F, G \in \mathcal{F}_r$ , the Mallows metric  $d_r(F, G)$  is defined by*

$$d_r(F, G) = \inf_{\mathcal{T}_{X,Y}} \{ \mathbb{E}|X - Y|^r \}^{1/r},$$

*where  $\mathcal{T}_{X,Y}$  is the set of all joint distributions of pairs of random variables  $X$  and  $Y$  whose marginal distributions are  $F$  and  $G$  respectively. We also write  $d_r(X, Y)$  for  $d_r(F, G)$ , where this will cause no confusion.*

The empirical process is a fundamental quantity of interest. Though often not explicitly recognised as such, the Mallows distance between the empirical distribution and the underlying population has arisen in the work of several authors, including Csörgő and Horváth (1990) and del Barrio et al. (1999, 2000).

Suppose  $X_1, \dots, X_n$  are independent random variables, each having distribution function  $F$  with mean  $\mu$  and finite variance  $\sigma^2 > 0$ , and let  $\hat{F}_n$  denote the empirical distribution function of the sample, given by

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}.$$

The main contributions of this paper are threefold:

1. In Theorem 2.6, we show that in the case of a discrete underlying population of finite support,  $n^{1/4}d_2(\hat{F}_n, F)$  converges to an explicit non-degenerate limiting distribution. We contrast this with the  $n^{1/2}$  normalisation required by the previously cited authors in the absolutely continuous case.

2. In Section 3, we study the tail conditions required by Csörgő and Horváth (1990) and del Barrio et al. (2000) for the convergence of  $n^{1/2}d_2(\hat{F}_n, F)$  in the absolutely continuous case. In particular, by considering the hazard function, we show that one of the conditions of del Barrio et al. (2000) is redundant, and the statement of Theorem 2.1 of Csörgő and Horváth (1990) may be simplified.
3. Section 4 is devoted to an application of these results. We recall the calculation of Shao and Tu (1995), showing that  $d_2(\hat{F}_n, F)$  provides an upper bound on the Mallows distance between the bootstrap distribution of the sample mean and its true sampling distribution. We give a straightforward lower bound on this latter quantity, yielding conditions under which the upper and lower bounds are of the same order.

## 2 Convergence rates and limiting distributions for $d_r(\hat{F}_n, F)$

First, recall the following two lemmas about  $d_r$ , which are proved in Major (1978) and Bickel and Freedman (1981) respectively.

**Lemma 2.1** *For  $F, G \in \mathcal{F}_r$ , the infimum in Definition 1.1 is attained by the following construction: let  $U \sim U(0, 1)$ , and set  $X = F^{-1}(U)$ ,  $Y = G^{-1}(U)$ , where, for example,  $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$ . Thus*

$$d_r(F, G) = \left( \int_0^1 |F^{-1}(p) - G^{-1}(p)|^r dp \right)^{1/r}.$$

**Lemma 2.2** *If  $(F_n) \in \mathcal{F}$  and  $F \in \mathcal{F}$ , then  $d_r(F_n, F) \rightarrow 0$  as  $n \rightarrow \infty$  if and only if, for every bounded, continuous function  $g : \mathbb{R} \rightarrow \mathbb{R}$ , we have both*

1.  $\lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} g(x) dF_n(x) = \int_{-\infty}^{\infty} g(x) dF(x);$
2.  $\lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} |x|^r dF_n(x) = \int_{-\infty}^{\infty} |x|^r dF(x).$

*Thus, convergence in the Mallows metric  $d_r$  is equivalent to convergence in distribution together with convergence of the  $r$ th absolute moments.*

It follows immediately by Lemma 2.2 and the strong law of large numbers that  $d_2(\hat{F}_n, F) \rightarrow 0$  almost surely as  $n \rightarrow \infty$ . It is important to remark that although we can calculate a rate of convergence of the two parts above (that is, convergence in distribution and of the  $r$ th absolute moment), this will not help us find a rate of convergence of  $d_2(\hat{F}_n, F)$ , and we must use other techniques.

When  $F$  has a density  $f$ , Csörgő and Horváth (1990) and del Barrio et al. (2000) give conditions under which  $n^{1/2}d_2(\hat{F}_n, F)$  converges to a nondegenerate limiting distribution. A simple version of such results is given in Theorem 2.4 below. However, these results do not cover the case of a discrete underlying population, which is studied later, in Theorem 2.6.

In order to prove Theorem 2.4, we need to introduce some notation. Let  $D = D[0, 1]$  denote the space of left-continuous, real-valued functions on  $[0, 1]$  possessing right limits at each point. We may equip  $D$  with the uniform norm

$$\|x - y\|_\infty = \sup_{p \in [0, 1]} |x(p) - y(p)|.$$

A small complication arises from the fact that the normed space  $(D, \|\cdot\|_\infty)$  is non-separable, and the  $\sigma$ -algebra,  $\mathcal{D}$ , generated by the open balls is strictly smaller than the Borel  $\sigma$ -algebra,  $\mathcal{D}_{\text{Borel}}$ , generated by the open sets. This creates measurability problems, as explained in Chibisov (1965), which lead us to work with the space  $(D, \mathcal{D}, \|\cdot\|_\infty)$ . A consequence of using the ball  $\sigma$ -algebra is that we must make a slight modification to the notion of weak convergence, in line with Billingsley (1999), p. 67:

**Definition 2.3** *If  $(Y_n)_{n \geq 0}$  is a sequence of random elements of  $(D, \mathcal{D}, \|\cdot\|_\infty)$ , we write  $Y_n \xrightarrow{d^\circ} Y_0$  as  $n \rightarrow \infty$  if*

$$\mathbb{E}(f(Y_n)) \rightarrow \mathbb{E}(f(Y_0))$$

*as  $n \rightarrow \infty$ , for all bounded, continuous functions  $f : D \rightarrow \mathbb{R}$  which are  $\mathcal{D}$ -measurable.*

Throughout,  $B = (B(p))_{0 \leq p \leq 1}$  denotes a Brownian bridge; that is, a zero mean Gaussian process with

$$\text{Cov}(B(p), B(q)) = p(1 - q)$$

for  $p \leq q$ . For  $p \in (0, 1)$ , let  $\xi_p = F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$ .

**Theorem 2.4** *Suppose that  $F$  has a density  $f$  such that  $f(\xi_p)$  is positive and continuous for  $p \in [0, 1]$ , and that the limits  $\xi_0 = \lim_{p \searrow 0} \xi_p$  and  $\xi_1 = \lim_{p \nearrow 1} \xi_p$  exist in  $\mathbb{R}$ . Then*

$$n^{1/2} d_2(\hat{F}_n, F) \xrightarrow{d} \left( \int_0^1 B(p)^2 (F^{-1})'(p)^2 dp \right)^{1/2}$$

as  $n \rightarrow \infty$ .

**Proof** Theorem 1 on pp. 640–641 of Shorack and Wellner (1986), together with Corollary 1 on p. 48 of the same book, give that

$$f(\xi_p) n^{1/2} (\hat{F}_n^{-1}(p) - \xi_p) \xrightarrow{d^\circ} B(p) \quad (1)$$

on  $(D, \mathcal{D}, \|\cdot\|_\infty)$ , as  $n \rightarrow \infty$ . Now, with probability one,  $B$  belongs to the space  $(C[0, 1], \|\cdot\|_\infty)$  of continuous real-valued functions on  $[0, 1]$  equipped with the uniform norm, and moreover this space is separable. We can therefore apply the version of the continuous mapping theorem for  $\xrightarrow{d^\circ}$  convergence (Billingsley, 1999, pp. 67–68) to a composition map  $h(p) = h_2(h_1(p))$  from  $(D, \mathcal{D}, \|\cdot\|_\infty)$  to  $\mathbb{R}$ . The individual maps  $h_1 : (D, \mathcal{D}, \|\cdot\|_\infty) \rightarrow (D, \mathcal{D}, \|\cdot\|_\infty)$  and  $h_2 : (D, \mathcal{D}, \|\cdot\|_\infty) \rightarrow \mathbb{R}$  are defined by

$$h_1(G)(p) = \frac{G(p)^2}{f(\xi_p)^2} \quad \text{and} \quad h_2(G) = \left( \int_0^1 G(p) dp \right)^{1/2}.$$

Observe that the continuity of  $h_1$  follows from the fact that  $f(\xi_p)$  attains its (positive) infimum for some  $p \in [0, 1]$ . We conclude that

$$n^{1/2} d_2(\hat{F}_n, F) = \left( \int_0^1 n (\hat{F}_n^{-1}(p) - \xi_p)^2 dp \right)^{1/2} \xrightarrow{d^\circ} \left( \int_0^1 \frac{B(p)^2}{f(\xi_p)^2} dp \right)^{1/2}$$

as  $n \rightarrow \infty$ . The result follows on noting that any bounded, continuous function from  $\mathbb{R}$  to  $\mathbb{R}$  is (Borel) measurable.  $\square$

Theorem 2.4 is not the strongest possible; in particular, it is not necessary for the underlying density to have bounded support. Csörgő and Horváth (1990) and del Barrio et al. (2000) obtain the same conclusion with conditions which amount to control of the behaviour of the tails of  $F$ . For example, Equations (3.16) and (3.14) of del Barrio et al. (2000) consist of the following two conditions:

**Condition 1** As  $n \rightarrow \infty$ ,

$$n \int_0^{1/n} (\hat{F}_n^{-1}(p) - F^{-1}(p))^2 \xrightarrow{p} 0 \quad \text{and} \quad n \int_{(n-1)/n}^1 (\hat{F}_n^{-1}(p) - F^{-1}(p))^2 dp \xrightarrow{p} 0.$$

**Condition 2**

$$\int_0^1 \frac{p(1-p)}{f(F^{-1}(p))^2} dp = \int_{-\infty}^{\infty} \frac{F(x)(1-F(x))}{f(x)} dx < \infty.$$

A version of Theorem 2.1 of Csörgő and Horváth (1990) gives the following:

**Theorem 2.5** Suppose that  $F$  has a density  $f$  such that  $f(\xi_p)$  is positive and continuous for  $p \in (0, 1)$ , and monotone for  $p$  sufficiently close to zero and one. If Conditions 1 and 2 hold, then

$$n^{1/2} d_2(\hat{F}_n, F) \xrightarrow{d} \left( \int_0^1 B(p)^2 (F^{-1})'(p)^2 dp \right)^{1/2}$$

as  $n \rightarrow \infty$ .

Proposition 1 of del Barrio et al. (1999) verifies Condition 1 when  $F$  is the normal distribution. We generalise this proposition in Section 3, by showing that Condition 1 holds when the hazard function diverges in the tails. In particular then, Condition 1 will be seen to be a consequence of Condition 2.

Now, we turn to the case of discrete random variables. Note that in Theorem 2.6 below, we can understand the limiting distribution as the integral between 0 and 1 of a sum of delta functions at the points where  $F^{-1}(\cdot)$  jumps. Thinking of the limit in this way shows the analogy with the limits in Theorem 2.4 and Theorem 2.5.

**Theorem 2.6** Let  $F$  be the distribution function corresponding to the probability mass function given by  $\mathbb{P}(X_1 = x_j) = p_j$ , where  $x_1 < x_2 < \dots < x_m$  and  $\sum_{j=1}^m p_j = 1$ , with each  $p_j > 0$ . For  $j = 1, \dots, m$ , let  $q_j = p_1 + \dots + p_j$ . Then

$$n^{1/4} d_2(\hat{F}_n, F) \xrightarrow{d} \left( \sum_{j=1}^{m-1} |B(q_j)|(x_{j+1} - x_j)^2 \right)^{1/2}$$

as  $n \rightarrow \infty$ .

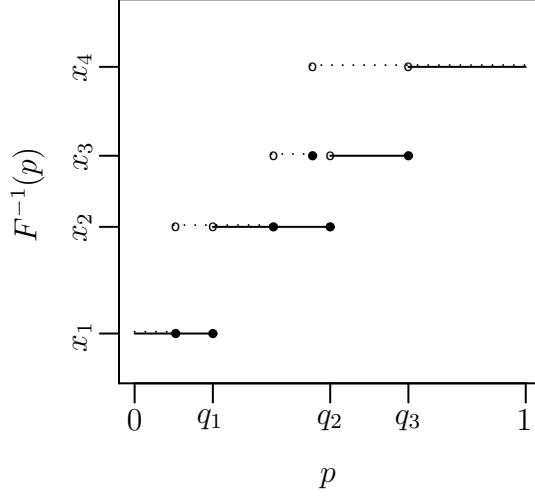


Figure 1: Graphs of  $F^{-1}(p)$  (solid) and  $\hat{F}_n^{-1}(p)$  (dotted). The dotted lines have been perturbed slightly to distinguish them from the solid ones.

**Proof** See Figure 1. For  $j = 1, \dots, m$ , define  $\hat{p}_j = n^{-1} \sum_{i=1}^n \mathbb{1}_{\{X_i = x_j\}}$  to be the empirical proportion of the sample taking the value  $x_j$ . Let  $\epsilon = \min_j p_j/3$ , write  $\hat{q}_j = \hat{p}_1 + \dots + \hat{p}_j$ , and define the event

$$A = \{|\hat{q}_j - q_j| \leq \epsilon \text{ for } j = 1, \dots, m\}.$$

Observe that, by the Dvoretzky–Keifer–Wolfowitz inequality (Massart, 1990),

$$\mathbb{P}(A^c) = \mathbb{P}\left(\bigcup_{j=1}^m |\hat{q}_j - q_j| > \epsilon\right) = \mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Thus, since  $|\hat{F}_n^{-1}(p) - F^{-1}(p)| \leq x_m - x_1$ ,

$$d_2^2(\hat{F}_n, F) \mathbb{1}_{A^c} \leq (x_m - x_1)^2 \mathbb{1}_{A^c} = o_p(n^{-1/2})$$

as  $n \rightarrow \infty$ . On the other hand,

$$d_2^2(\hat{F}_n, F) \mathbb{1}_A = \sum_{j=1}^{m-1} |\hat{q}_j - q_j| (x_{j+1} - x_j)^2.$$

But  $(n^{1/2}(\hat{q}_1 - q_1), \dots, n^{1/2}(\hat{q}_m - q_m)) \xrightarrow{d} N_m(0, \Sigma)$ , where the asymptotic covariance matrix  $\Sigma$  has entries  $\Sigma_{ij} = q_i(1 - q_j)$ , for  $1 \leq i \leq j \leq m$ . So by the continuous mapping theorem,

$$n^{1/2} \sum_{j=1}^{m-1} |\hat{q}_j - q_j| (x_{j+1} - x_j)^2 \xrightarrow{d} \sum_{j=1}^{m-1} |B(q_j)| (x_{j+1} - x_j)^2$$

as  $n \rightarrow \infty$ . Hence

$$n^{1/2} d_2^2(\hat{F}_n, F) = n^{1/2} d_2^2(\hat{F}_n, F) \mathbb{1}_A + o_p(1) \xrightarrow{d} \sum_{j=1}^{m-1} |B(q_j)| (x_{j+1} - x_j)^2$$

as  $n \rightarrow \infty$ . □

Theorems 2.4 and 2.6 extend to the case of general  $d_r$ . Note the rates below coincide only when  $r = 1$ , a case studied by del Barrio et al. (1999b).

### Corollary 2.7

(a) Suppose that  $F$  has a density  $f$  such that  $f(\xi_p)$  is positive and continuous for  $p \in [0, 1]$ , and that the limits  $\xi_0 = \lim_{p \searrow 0} \xi_p$  and  $\xi_1 = \lim_{p \nearrow 1} \xi_p$  exist in  $\mathbb{R}$ . Then

$$n^{1/2} d_r(\hat{F}_n, F) \xrightarrow{d} \left( \int_0^1 |B(p)|^r (F^{-1})'(p)^r dp \right)^{1/2}$$

as  $n \rightarrow \infty$ .

(b) Let  $F$  be the distribution function corresponding to the probability mass function given by  $\mathbb{P}(X_1 = x_j) = p_j$ , where  $x_1 < x_2 < \dots < x_m$  and  $\sum_{j=1}^m p_j = 1$ , with each  $p_j > 0$ . Let  $q_j = p_1 + \dots + p_j$ . Then

$$n^{1/(2r)} d_r(\hat{F}_n, F) \xrightarrow{d} \left( \sum_{j=1}^{m-1} |B(q_j)| (x_{j+1} - x_j)^r \right)^{1/r}$$

as  $n \rightarrow \infty$ .



**Proof** The proof of part (a) follows that of Theorem 2.4, the only difference being that the function  $h_1$  should in this case be defined by

$$h_1(G)(p) = \frac{|G(p)|^r}{f(\xi_p)^r}.$$

Part (b) mirrors the proof of Theorem 2.6, noting that in general,

$$n^{1/2} d_r^r(\hat{F}_n, F) \mathbb{1}_A = n^{1/2} \sum_{j=1}^{m-1} |\hat{q}_j - q_j| (x_{j+1} - x_j)^r,$$

where the event  $A$  was defined in the proof of Theorem 2.6.  $\square$

### 3 The hazard function and tail behaviour

In Section 2, we saw the importance of Conditions 1 and 2 concerning the tail behaviour of the distribution function  $F$  in determining the rate of convergence to zero of  $d_2(\hat{F}_n, F)$ . We now study the nature of these conditions in greater detail, using the notion of the hazard function.

**Definition 3.1** *For a random variable  $X$  with distribution function  $F$  and continuous density  $f$ , we define  $\bar{F}(x) = \mathbb{P}(X > x)$ , and the two-sided hazard function by*

$$h(x) = \begin{cases} f(x)/\bar{F}(x) & \text{if } x \geq \xi_{1/2}, \\ f(x)/F(x) & \text{if } x < \xi_{1/2}. \end{cases}$$

The following theorem is the main result of this section, showing the redundancy of Condition 1. Its proof is deferred until after Lemma 3.4.

**Theorem 3.2** *If the hazard function satisfies  $h(x) \rightarrow \infty$  as  $|x| \rightarrow \infty$ , then Condition 1 holds; that is,*

$$n \int_0^{1/n} (\hat{F}_n^{-1}(p) - F^{-1}(p))^2 dp \xrightarrow{p} 0 \quad \text{and} \quad n \int_{(n-1)/n}^1 (\hat{F}_n^{-1}(p) - F^{-1}(p))^2 dp \xrightarrow{p} 0$$

*as  $n \rightarrow \infty$ . Moreover, Condition 2 implies Condition 1.*

Let  $X_t$  denote a random variable distributed as  $X|X > t$ , and consider  $\text{Var } X_t$  (we set  $\text{Var } X_t = 0$  if  $\mathbb{P}(X > t) = 0$ ). For example, if  $X$  is an exponential random variable, then  $X_t \sim X + t$ , and we have  $\text{Var } X_t = \text{Var } X$ . In general, this tail variance is closely related to the hazard function, as shown by the following lemma.

**Lemma 3.3** *Let  $X$  be a random variable with hazard function  $h(x)$ . If  $t \geq \xi_{1/2}$ , then*

$$\frac{1}{12 \sup_{x \geq t} h(x)^2} \leq \text{Var } X_t \leq \frac{4}{\inf_{x \geq t} h(x)^2},$$

*with a corresponding result for the left-hand tail.*

**Proof**

**(a) Proof of the lower bound.** Define the function

$$k(t) = \frac{\sup_{x \geq t} f(x)}{\bar{F}(t)}.$$

Note that the density of  $X_t$  for  $x \geq t$  is  $f(x)/\bar{F}(t)$ , which is at most  $k(t)$ . By comparison with the  $U[t, t + 1/c]$  distribution, we see that if a random variable has density bounded above by  $c$ , then its variance is at least  $1/(12c^2)$ , so  $\text{Var } X_t \geq 1/(12k(t)^2)$ . Moreover, if  $x \geq t$ , then  $f(x)/\bar{F}(t) \leq f(x)/\bar{F}(x)$ , from which we deduce that  $k(t) \leq \sup_{x \geq t} h(x)$ .

**(b) Proof of the upper bound.** In fact, we establish the stronger conclusion that

$$\frac{\mathbb{E}((X - t)^2 \mathbb{1}_{\{X > t\}})}{\bar{F}(t)} \leq \frac{4}{\inf_{x \geq t} h(x)^2}. \quad (2)$$

We use arguments based on those which establish Poincaré inequalities, in Borovkov and Utev (1984) and Sysoeva (1965). The simplest possible case of the main theorem of Sysoeva (1965) gives a Hardy inequality, that if  $G$  is a differentiable function with  $G(t) = 0$ , then

$$\int_t^\infty f(x) G(x)^2 dx \leq 4 \int_t^\infty \frac{\bar{F}(x)^2}{f(x)} g(x)^2 dx = 4 \int_t^\infty f(x) \frac{g(x)^2}{h(x)^2} dx, \quad (3)$$

where  $g(x) = G'(x)$ , provided the integrals exist.

[For the sake of completeness, note that Equation (3) can be proved as follows. Integration by parts gives that

$$\begin{aligned} \int_t^\infty f(x)G(x)^2 dx &= 2 \int_t^\infty \bar{F}(x)G(x)g(x) dx \\ &\leq 2 \left( \int_t^\infty f(x)G(x)^2 dx \right)^{1/2} \left( \int_t^\infty \frac{\bar{F}(x)^2}{f(x)} g(x)^2 dx \right)^{1/2}, \end{aligned}$$

by Cauchy-Schwarz.] Hence, from Equation (3),

$$\int_t^\infty f(x)G(x)^2 dx \leq \frac{4}{\inf_{x \geq t} h(x)^2} \int_t^\infty f(x)g(x)^2 dx,$$

and choosing  $G(x) = x - t$ , Equation (2) follows.  $\square$

This shows that if  $h(t)$  is bounded away from zero for large  $|t|$ , then  $\text{Var } X_t < \infty$ . This is the case for the normal distribution, where the Mills ratio (Shorack and Wellner, 1986, p. 850) gives that  $h(t) \geq |t|$  for all  $t$ . On the other hand, if  $h(t) \rightarrow 0$  as  $|t| \rightarrow \infty$ , then  $\text{Var } X_t \rightarrow \infty$ , as for the log-normal distribution, where  $h(t) \sim t^{-1} \log t$  as  $t \rightarrow \infty$ . Finally we remark that other tail variance behaviour is possible, in the case where  $\liminf_{t \rightarrow \infty} h(t) < \limsup_{t \rightarrow \infty} h(t)$ .

As the example of the log-normal might suggest, results from reliability theory provide a link between the hazard function and the finiteness of the moment generating function in a neighbourhood of the origin.

**Lemma 3.4** *Let  $X$  be a random variable with hazard function  $h(x)$ . If, for some  $t \geq 0$ , we have  $\inf_{|x| \geq t} h(x) \geq c$ , then  $\mathbb{E}(e^{\theta X}) < \infty$  for  $|\theta| < c$ .*

**Proof** We prove the result for the right-hand tail. Observe that for  $x \geq t$ ,

$$\log \left( \frac{\bar{F}(x)}{\bar{F}(t)} \right) = - \int_t^x h(y) dy,$$

which can be seen by differentiating both sides with respect to  $x$ . Thus

$$\bar{F}(x) = \bar{F}(t) \exp \left( - \int_t^x h(y) dy \right). \quad (4)$$

Now the right-hand side of Equation (4) is bounded above by  $e^{-cx}$ , which is enough to guarantee that the moment generating function is finite for  $\theta \in [0, c)$ .  $\square$

Lemma 3.4 shows that if a random variable  $X$  has a moment generating function which is infinite other than at the origin, then there exists a sequence  $(x_n)$  tending to infinity such that  $\lim_{n \rightarrow \infty} h(x_n) = 0$ . If, further,  $X$  has a decreasing hazard function then we deduce that  $\lim_{x \rightarrow \infty} h(x) = 0$ , and hence by Lemma 3.3, that  $\lim_{t \rightarrow \infty} \text{Var } X_t = \infty$ .

We are now in a position to prove Theorem 3.2.

**Proof of Theorem 3.2** We have

$$\begin{aligned}
n \int_{(n-1)/n}^1 (\hat{F}_n^{-1}(p) - F^{-1}(p))^2 dp \\
&= n \int_{(n-1)/n}^1 (\hat{F}_n^{-1}(p) - \xi_{(n-1)/n} + \xi_{(n-1)/n} - F^{-1}(p))^2 dp \\
&\leq 2 \left( (X_{(n)} - \xi_{(n-1)/n})^2 + n \int_{\xi_{(n-1)/n}}^{\infty} (x - \xi_{(n-1)/n})^2 dF(x) \right).
\end{aligned} \tag{5}$$

The second term tends to zero by Equation (2). If  $\xi_1 < \infty$ , then the first term clearly converges in probability to zero. On the other hand, if  $\xi_1 = \infty$ , then the assumed hypothesis on the hazard function combined with Equation (4) imply that for any  $t > 0$ ,

$$\lim_{x \rightarrow \infty} \frac{\bar{F}(t+x)}{\bar{F}(x)} = 0.$$

These are precisely the conditions under which Galambos (1978), Theorem 4.1.2, proves convergence in probability of  $X_{(n)} - \xi_{(n-1)/n}$  to zero. This completes the first part of the proof.

Note that Condition 2 can be restated in terms of the hazard function; that is, by considering the regions  $x \in [\xi_{1/2}, \infty)$  and  $x \in (-\infty, \xi_{1/2})$  separately, we see that Condition 2 holds if and only if

$$\int_{-\infty}^{\infty} \frac{1}{h(x)} dx < \infty.$$

Finiteness of this integral implies that  $h(x) \rightarrow \infty$  as  $|x| \rightarrow \infty$ , and so by the argument above, Condition 1 holds.  $\square$

Note for the normal distribution that  $h(x) \rightarrow \infty$ , but, again by the Mills ratio,

$$\int_1^\infty \frac{1}{h(x)} dx \geq \int_1^\infty \left( \frac{1}{x} - \frac{1}{x^3} \right) dx = \infty.$$

Thus Condition 1 holds (implying Proposition 1 of del Barrio et al. (1999)), but not Condition 2.

Finally in this section, we give a partial converse to Theorem 2.5.

**Proposition 3.5** *If  $\text{Var } X_t \rightarrow \infty$  as  $t \rightarrow \infty$ , then  $n^{1/2}d_2(\hat{F}_n, F) \xrightarrow{a.s.} \infty$  as  $n \rightarrow \infty$ .*

**Proof** For  $i = 1, \dots, n$ , let  $a_i = n \int_{(i-1)/n}^{i/n} F^{-1}(p) dp$ . We decompose  $nd_2^2(\hat{F}_n, F)$  into a random and deterministic part as follows:

$$\begin{aligned} nd_2^2(\hat{F}_n, F) &= n \sum_{i=1}^n \int_{(i-1)/n}^{i/n} ((X_{(i)} - a_i)^2 + (F^{-1}(p) - a_i)^2) dp \\ &= \sum_{i=1}^n (X_{(i)} - a_i)^2 + n \sum_{i=1}^n \int_{(i-1)/n}^{i/n} (F^{-1}(p) - a_i)^2 dp \\ &\geq (X_{(n)} - a_n)^2 + n \int_{(n-1)/n}^1 (F^{-1}(p) - a_n)^2 dp \\ &= (X_{(n)} - a_n)^2 + n \int_{\xi_{(n-1)/n}}^\infty (x - a_n)^2 dF(x) \\ &= (X_{(n)} - a_n)^2 + \text{Var } X_{\xi_{(n-1)/n}}. \end{aligned}$$

Ignoring the random part, the result is immediate.  $\square$

## 4 Application to the bootstrap

The bootstrap was introduced into Statistics in the landmark paper of Efron (1979), and gives a very general technique for approximating the distributions

of roots (i.e. functions of the sample and parameters of the underlying population) of interest. The key idea for its use in practice is that of resampling; given a sample  $X_1, \dots, X_n$ , we draw a further sample  $X_1^*, \dots, X_n^*$  uniformly at random with replacement from the original sample, and perform our calculations based on this resample. The opportunity to repeat this resampling procedure enables the practitioner to mimic drawing additional samples from the original population.

Historically, one of the great early triumphs in the analysis the bootstrap was the paper of Singh (1981). One of his main results was to show that the bootstrap distribution of a normalised sample mean converges to its true sampling distribution at rate  $O(n^{-1})$ , provided the underlying population is non-lattice, and sufficiently many moments exist. This improves on the  $O(n^{-1/2})$  convergence rate of its natural competitor, namely normal approximation. Singh's theorem was stated in terms of the supremum distance between the respective distribution functions. The Mallows metric was first considered in the context of the bootstrap by Bickel and Freedman (1981).

Recall that  $X_1, \dots, X_n$  are independent random variables, each with distribution function  $F$  with mean  $\mu$  and finite variance  $\sigma^2 > 0$ , that  $\hat{F}_n$  is the empirical distribution function of the sample and that  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$  is the sample mean. A standard procedure for constructing confidence intervals for  $\mu$  is to invert a probability statement concerning a root such as  $n^{1/2}(\bar{X}_n - \mu)$ , whose sampling distribution under  $F$  we denote by  $H_n(F)$ . Conditional on  $X_1, \dots, X_n$ , let  $X_1^*, \dots, X_n^*$  be a resample; that is, an independent and identically distributed sample drawn from  $\hat{F}_n$ .

The nonparametric bootstrap estimates the sampling distribution of  $n^{1/2}(\bar{X}_n - \mu)$  by that of  $n^{1/2}(\bar{X}_n^* - \bar{X}_n)$ , where  $\bar{X}_n^* = n^{-1} \sum_{i=1}^n X_i^*$ . In other words, conditional on  $X_1, \dots, X_n$ , we approximate  $H_n(F)$  by  $H_n(\hat{F}_n)$ . The properties of the Mallows distance  $d_2$  make it suitable for studying the performance of the bootstrap approximation in this context. The calculation below follows

Shao and Tu (1995), and uses results proved in Bickel and Freedman (1981).

$$\begin{aligned}
d_2(H_n(\hat{F}_n), H_n(F)) &= d_2\left(\frac{1}{n^{1/2}} \sum_{i=1}^n (X_i^* - \bar{X}_n), \frac{1}{n^{1/2}} \sum_{i=1}^n (X_i - \mu)\right) \\
&\leq \frac{1}{n^{1/2}} \left( \sum_{i=1}^n d_2(X_i^* - \bar{X}_n, X_i - \mu)^2 \right)^{1/2} \\
&= d_2(X_1^* - \bar{X}_n, X_1 - \mu) \\
&= (d_2(X_1^*, X_1)^2 - (\bar{X}_n - \mu)^2)^{1/2} \\
&\leq d_2(X_1^*, X_1) \\
&= d_2(\hat{F}_n, F).
\end{aligned} \tag{6}$$

Thus, in particular, the distance between the distribution of the root of interest,  $H_n(F)$ , and its bootstrap approximation,  $H_n(\hat{F}_n)$ , is stochastically dominated by the distance between the true and empirical distributions. The first inequality and the conditions under which we obtain equality are studied in Johnson and Samworth (2004).

Having studied the upper bound  $d_2(\hat{F}_n, F)$  in previous sections, our final result gives a lower bound on the rate of convergence of  $d_2(H_n(\hat{F}_n), H_n(F))$ . Observe in particular that if the hypotheses of Theorem 2.4 and Proposition 4.1 are both satisfied, then the rates of convergence of the two bounds are identical.

#### Proposition 4.1

- (a) *If  $\mathbb{E}X_1^4 < \infty$ , then  $n^{1/2}d_2(H_n(\hat{F}_n), H_n(F))$  is bounded away from zero in probability; in fact, for every  $\epsilon > 0$ ,*

$$\mathbb{P}\{n^{1/2}d_2(H_n(\hat{F}_n), H_n(F)) > \epsilon\} \rightarrow 0$$

*as  $n \rightarrow \infty$ .*

- (b) *Fix  $\delta \in (0, 2)$ . If  $\mathbb{E}|X_1|^{2+\delta} = \infty$ , then, with probability one,  $n^{\delta/(2+\delta)}d_2(H_n(\hat{F}_n), H_n(F))$  does not tend to zero; that is*

$$\mathbb{P}\{n^{\delta/(2+\delta)}d_2(H_n(\hat{F}_n), H_n(F)) \rightarrow 0\} = 0.$$

**Proof** For any  $r \geq 1$  and  $F_X, F_Y \in \mathcal{F}_r$ , there exist random variables  $(X, Y)$  having marginal distribution functions  $F_X$  and  $F_Y$  respectively, and such that

$$d_r(F_X, F_Y) = (\mathbb{E}|X - Y|^r)^{1/r} \geq |(\mathbb{E}|X|^r)^{1/r} - (\mathbb{E}|Y|^r)^{1/r}|. \quad (7)$$

This means that

$$\begin{aligned} d_2(H_n(\hat{F}_n), H_n(F)) &= d_2\left(\frac{1}{n^{1/2}} \sum_{i=1}^n (X_i^* - \bar{X}_n), \frac{1}{n^{1/2}} \sum_{i=1}^n (X_i - \mu)\right) \\ &\geq |s - \sigma| \end{aligned}$$

where  $s^2 = n^{-1} \sum (X_i - \bar{X}_n)^2$  is the sample variance. Without loss of generality, suppose that  $\mu = \mathbb{E}X_1 = 0$ .

To prove (a), observe that if  $\mathbb{E}(X_1^4) < \infty$ , then  $n^{1/2}|s - \sigma|$  is bounded away from zero in probability; in fact,  $n^{1/2}(s - \sigma) \xrightarrow{d} N(0, (\mathbb{E}X_1^4 - \sigma^4)/(4\sigma^2))$  (Serfling, 1980, p. 119).

To prove (b), note that

$$\begin{aligned} n^{\delta/(2+\delta)}(s^2 - \sigma^2) &= \frac{1}{n^{2/(2+\delta)}} \sum_{i=1}^n \{(X_i - \bar{X}_n)^2 - \sigma^2\} \\ &= \frac{1}{n^{2/(2+\delta)}} \sum_{i=1}^n (X_i^2 - \sigma^2) - n^{\delta/(2+\delta)} \bar{X}_n^2. \end{aligned}$$

Now, applying the law of the iterated logarithm gives that

$$n^{\delta/(2+\delta)} \bar{X}_n^2 = O(n^{-2/(2+\delta)} \log \log n) \quad \text{a.s..}$$

Moreover, since  $\mathbb{E}|X_1|^{2+\delta} = \infty$ , by the converse to the Marcinkiewicz-Zygmund strong law of large numbers (cf. Loève (1977), p. 255),

$$\mathbb{P}\left(\frac{1}{n^{2/(2+\delta)}} \sum_{i=1}^n (X_i^2 - \sigma^2) \rightarrow 0\right) < 1.$$

Hence, by Kolmogorov's zero-one law, and since  $s^2 - \sigma^2 = (s - \sigma)(s + \sigma)$ ,

$$\mathbb{P}(n^{\delta/(2+\delta)}|s - \sigma| \rightarrow 0) = 0.$$

□

Note that the strength of the conclusion of part (b) of Proposition 4.1 increases as  $\delta$  decreases.



## References

- Bickel, P. J. and Freedman, D. A. (1981), *Some asymptotic theory for the bootstrap*, Ann. Statist., **9**, 1196–1217.
- Billingsley, P. (1999), *Convergence of Probability Measures*, Second ed., Wiley, New York.
- Borovkov, A. and Utev, S. (1984), *On an inequality and a related characterisation of the normal distribution*, Theor. Probab. Appl., **28**, 219–228.
- Chibisov, D. M. (1965), *An investigation of the asymptotic power of the tests of fit*, Theor. Probab. Appl., **10**, 421–437.
- Csörgő, M. and Horváth, L. (1990), *On the distributions of  $L_p$  norms of weighted quantile processes*, Ann. Inst. Henri Poincaré, **26**, 65–90.
- del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. and Rodríguez-Rodríguez, J. M. (1999), *Tests of goodness of fit based on the  $L_2$ -Wasserstein distance*, Ann. Statist., **27**, 1230–1239.
- del Barrio, E., Giné, E. and Matrán, C. (1999b), *Central limit theorems for the Wasserstein distance between the empirical and the true distributions*, Ann. Probab., **27**, 1009–1071.
- del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. (2000), *Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests*, Trabajos Investigación Oper., **9**, 1–96.
- Efron, B. (1979), *Bootstrap methods: another look at the jackknife*, Ann. Statist., **7**, 1–26.
- Fisz, M. (1963), *Probability Theory and Mathematical Statistics*, Third ed., Wiley, New York.
- Galambos, J. (1978), *The Asymptotic Theory of Extreme Order Statistics*, Wiley, New York.
- Johnson, O. T. and Samworth, R. J. (2004), *Central limit theorem and convergence to stable laws in Mallows distance*, submitted to Bernoulli and available at tex-arkiv paper no. math.PR/0406218.

- Loève, M. (1977), *Probability Theory I*, Fourth Ed., Springer-Verlag, New York.
- Major, P. (1978), *On the invariance principle for sums of independent, identically distributed random variables*, J. Mult. Anal., **8**, 487–501.
- Massart, P. (1990), *The tight constant in the Dvoretzky–Keifer–Wolfowitz inequality*, Ann. Probab., **18**, 1269–1283.
- Rachev, S. T. (1984), *The Monge-Kantorovich problem on mass transfer and its applications in stochastics*, Theor. Probab. Appl., **29**, 647–676.
- Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.
- Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag, New York.
- Shorack, G. R. and Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*, Wiley, New York.
- Singh, K. (1981), *On the asymptotic validity of Efron’s bootstrap*, Ann. Statist., **9**, 1187–1195.
- Sysoeva, F. A. (1965), *Generalization of a certain Hardy inequality*, Izv. Vysš. Učebn. Zaved. Matematika, **49**, 140–143.